

# Survival Model Construction Guided by Fit and Predictive Strength

Cécile Chauvel and John O’Quigley

Laboratoire de Statistique Théorique et Appliquée,  
Université Pierre et Marie Curie - Paris VI, 75005 Paris, France

**ABSTRACT:** We describe a unified framework within which we can build survival models. The motivation for this work comes from a study on the prediction of relapse among breast cancer patients treated at the Curie Institute in Paris, France. Our focus is on how to best code, or characterize, the effects of the variables, either alone or in combination with others. We consider simple graphical techniques that not only provide an immediate indication as to the goodness of fit but, in cases of departure from model assumptions, point in the direction of a more involved alternative model. These techniques help support our intuition. This intuition is backed up by formal theorems that underlie the process of building richer models from simpler ones. Goodness-of-fit techniques are used alongside measures of predictive strength and, again, formal theorems show that these measures can be used to help identify models closest to the unknown non-proportional hazards mechanism that we can suppose generates the observations. We consider many examples and show how these tools can be of help in guiding the practical problem of efficient model construction for survival data.

*Key words:* Proportional Hazards; Goodness of fit; Predictive measures;  $R^2$  measures; Time-varying coefficient.

## 1 INTRODUCTION

### 1.1 Motivation

The advent of personalized medicine together with rapid progress in techniques of genetics, next generation sequencing for example, the use of biomarkers, together with analytic techniques in bioinformatics have brought a renewed focus on the problems of model-based prediction. The related but different question concerning goodness of fit for any model has been given rather greater attention, at least in the survival literature.  $R^2$  measures quantify the predictive capacity of a model, and this may be high even when the model assumptions are seriously violated, whereas goodness-of-fit measures focus on the model assumptions and aim to examine how well these are supported by the data themselves. Although a number of authors have carefully outlined that distinction

it is true that some confusion still remains.

Our motivation stems from a study of 1504 breast cancer patients treated at the Institut Curie in Paris, France. Subsequent to initial treatment, patients were followed for a period of fifteen years. Among several study objective relating to this cohort was the aim to construct descriptive survival models that could provide a deeper understanding to prognosis after initial treatment. The problem is inherently a multi-factorial one. Combined effects of prognostic factors as well as conditional effects are a central concern. By conditional, we mean the impact of various risk factors on survival after having taken account of the impact of known or suspected risk factors. For instance, it can be of interest to try to quantify the added prognostic information of a more or less complex construction of biomarkers after having already accounted for known clinical risk factors. Finally, the effect of several of these risk factors can change with time and useful prognostic indices should reflect such time dependencies.

## 1.2 Background

Goodness-of-fit procedures can be directed at more than one aspect of any model. We may wish to consider overall fit of the model, i.e., how well the model when taken as a whole is supported by the observations or we may wish to focus on some particular feature of the model and how well it holds up in practice. For example, we may be interested in checking the working assumptions regarding treatment differences in presence of other covariates when the model fit of these covariates is of only indirect concern. The goodness of fit of a model can be evaluated by using tests or graphical methods. In this paper, we focus on graphical methods that can not only indicate departures from working assumptions but can also, of themselves, suggest remedies. The first graphical method for checking the proportional hazards assumption was proposed by Kay (1977) who suggested to plot an estimate of the conditional cumulative hazard  $\Lambda(t|Z)$  over time. When  $Z$  is a categorical covariate, typically representing treatment groups, the plot should result in parallel curves under proportional hazards. Andersen (1982) extended this approach to continuous covariates by discretizing them. Other graphical methods based on residuals can be sorted in two categories, depending on whether the residuals are cumulated or not. Amongst non-cumulative methods, a large class of martingale residuals described by Barlow and Prentice (1988) can be used by plotting their members over time. The Schoenfeld (1982) residuals, weighted Schoenfeld residuals introduced by Lin et al. (1993) and the residuals of Kay (1977) arise as special members of this class. Grambsch and Therneau (1994) suggested plotting standardized residuals over time to detect the validity of proportional hazards assumption and, in case of rejection, have an indication on the shape of the time-varying effect. This is the most commonly used approach and is implemented in the programming languages R and Splus. More recently, Sasieni and Winnett (2003) proposed the use of martingale difference residuals. The latter method requires care in interpretation since several plots

corresponding to several time points have to be considered. All of these non-cumulative residual methods presented so far make use of a smoothing function to average the residual points. As pointed out by Lin et al. (1993), the result can be sensitive to the choice of the smoothing techniques. To overcome this problem, several authors proposed the use of cumulative martingale residuals, such as Arjas (1988), Therneau et al. (1990) and Lin et al. (1993). The method of Therneau et al. (1990) is based on the score process of Wei (1984). Under the proportional hazards assumption, this process converges weakly to a Brownian bridge and a test of the supremum of a Brownian bridge can be performed. Lin et al. (1993) showed that Wei's score process can be asymptotically approximated by a gaussian process with a data-based variance-covariance matrix. Therefore, the comparison between the observed score process and a large numbers of simulated outcomes of the limiting gaussian process can give an indication of the validity of the proportional hazards assumption. In practice, the interpretation of such a plot is not always clear. More details about goodness-of-fit methods can be found in Klein and Moeschberger (2003), Therneau and Grambsch (2000) and more recently Martinussen and Scheike (2005).

Unlike the case of linear regression, if the multivariate proportional hazards model holds, the sub-models will no longer be simultaneously valid. Therefore, the evaluation of the goodness of fit of the multivariate model by evaluating the fit of the univariate sub-models will not suffice. However, in absence of tools for checking the overall validity of the model, most of the existing methods for checking the fit of one covariate assume proportional hazards for the other covariates, which is an erroneous assumption (Scheike and Martinussen 2004). Besides, the validity of the results of such methods depends on the covariance between covariates. To adress this issue, Scheike and Martinussen (2004) considered a non-proportional hazards model and developed estimation procedures and tests of the goodness of fit for one covariate with the possibility for the others not to have a constant regression effect. Their simulation work indicates the good performance of their method when compared to several existing and commonly used methods when the proportional hazards assumption is not met and/or in the presence of correlated covariates. Their test statistic depends on the estimation of the regression parameter requiring an involved algorithm relying on kernel estimation. The shape of the resulting estimator of the regression parameter is not an explicit and smooth function of time. The expression of the asymptotic distribution is unavailable for their statistic. The goodness-of-fit evaluation procedure presented in this article is a graphical method which does not require any estimation and is simple to understand. Our method is also based on the general framework of a non-proportional hazards model and is adapted to multivariate settings with correlated covariates.

Measures of predictive ability, on the other hand, - we will focus specifically on  $R^2$  type measures - are used to examine several different questions. Typical questions may be, how well does some set of biological markers perform, in a predictive sense, when

compared to some other set. How much added predictive information is contained in a biomarker when added to already known clinical prognostic factors such as stage and grade. When all known factors are included in a model, how much of the variability is accounted for so that, in consequence, how much variability remains to be explained, either by physical or possibly genetic attributes. Finally, how does the relaxing of certain model assumptions - one example would be stratification rather than inclusion in the linear component of a proportional hazards model - impact prediction. This last observation draws attention to the fact that, although different techniques with a different purpose, the aims of goodness-of-fit procedures and predictive measures can to some degree overlap. In the context of survival analysis, in particular when using the Cox proportional hazards model, several authors have proposed different measures of predictive ability. A recent and exhaustive literature review on the predictive accuracy measures can be found in Choodari-Oskooei et al. (2012). No consensus has yet been established regarding the most suitable measure to use in practice (Müller et al. 2008, Hielscher et al. 2010, Choodari-Oskooei et al. 2012).

It is not clear in what way, or in what sense, an improvement in predictability implies an improvement in goodness of fit. In fact it is not difficult to come up with counter examples and the notion itself is not very precise. The converse is however correct, and, in this work, we prove in a theorem that an improvement in fit of a proportional hazards model results in an improvement in predictability. This theorem underlies the purpose of this article which is to investigate ways to improve goodness-of-fit for proportional hazards type models and to see how this impacts the resulting predictive power of the model. We work with goodness-of-fit procedures and measures of predictive ability that are closely related, having as their basis the residuals from the non-proportional hazards model. The goodness of fit is evaluated with a version of the score process introduced by O’Quigley (2003, 2008 chap. 8) which is extended here to the multivariate setting. We obtain the exact expression of the limiting distribution of the process. The predictive accuracy measure is the  $R^2$  coefficient described by O’Quigley and Flandre (1994) but is also extended to the multivariate non-proportional hazards situation. This leads to easily assessed visual techniques and provides a complete and unified approach to the testing, fit and quantification of predictive effects. Several examples illustrate the ideas.

In the next section we describe the non-proportional hazards model and use it to derive stochastic processes of particular relevance to the problem we are studying. In Section 3, we present the main result that indicates why improvements in fit will result in improvements in predictive capability and how to proceed in practice. Section 4 summarizes simulations that provide additional support to our intuition and an application to a real dataset is provided. Before that, we recall the main notation.

### 1.3 Notation

The random variables of interest are the failure times  $T_i$ , the censoring times  $C_i$  and the vector of dimension  $p$  of possibly time-dependent covariates  $\mathbf{Z}_i = (Z_i^1, \dots, Z_i^p)$ ,  $i = 1, \dots, n$ . We view these as a random sample from the distribution of  $T$ ,  $C$  and  $\mathbf{Z} = (Z^1, \dots, Z^p)$  which have support on some finite interval. To emphasis the time-dependence, with a slight abuse of notation, we refer to any time-dependent quantity  $A$  as  $A(t)$ ,  $A$  being either random or deterministic. The time-dependent covariate  $\mathbf{Z}(t)$  is assumed to be a predictable stochastic process which admits a moment of order 4. For each subject  $i$ , the observed time is  $X_i = \min(T_i, C_i)$ , and the observed indicator of failure is  $\delta_i = I(T_i \leq C_i)$ , where  $I$  is the indicator function. The at-risk indicator  $Y_i(t)$  is defined as  $Y_i(t) = I(X_i \geq t)$ . The counting process  $N_i(t)$  is defined as  $N_i(t) = I(T_i \leq t, T_i \leq C_i)$  and we also define  $\bar{N}(t) = \sum_{i=1}^n N_i(t)$ . It is of notational convenience to define  $\mathcal{Z}(t) = \sum_{i=1}^n \mathbf{Z}_i(t) I(X_i = t, \delta_i = 1)$ , in words a  $\mathbb{R}^p$ -valued function equal to zero except at the observed failures where it assumes the covariate value of the subject that fails. In addition,  $\|\mathbf{a}\| = \max_{i=1, \dots, p} |a_i|$  denotes the maximum norm of the vector  $\mathbf{a} = (a_1, \dots, a_p) \in \mathbb{R}^p$ . For a  $p \times p$  matrix  $\mathbf{A}$  with element  $(i, j)$  denoted  $A_{i,j}$ ,  $i, j = 1, \dots, p$ ,  $\|\mathbf{A}\| = \max_{i,j=1, \dots, p} |A_{i,j}|$  denotes the maximum norm of  $\mathbf{A}$ . Let  $\mathbf{A}^T$  (respectively  $\mathbf{a}^T$ ) denote the transpose of the matrix  $\mathbf{A}$  (resp. vector  $\mathbf{a}$ ). The product  $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^T$  is the matrix with element  $[\mathbf{a}^{\otimes 2}]_{i,j} = a_i a_j$ . Denote  $\det(\mathbf{A})$  the determinant of the matrix  $\mathbf{A}$ . The space  $D[0, 1]^p = D[0, 1] \times \dots \times D[0, 1]$  is equipped with the Skorokhod product topology.

## 2 MODEL-BASED EMPIRICAL PROCESSES

Consider the non-proportional hazards model defined by

$$\lambda\{t \mid \mathbf{Z}(t)\} = \lambda_0(t) \exp\{\boldsymbol{\beta}(t)^T \mathbf{Z}(t)\}, \quad (1)$$

where  $\lambda(t|\cdot)$  is the conditional hazard function,  $\lambda_0(t)$  is a baseline hazard,  $\boldsymbol{\beta}(t)$  is the time-dependent regression effect and has dimension  $p$  and  $\boldsymbol{\beta}(t)^T \mathbf{Z}(t)$  is the usual inner product between  $\boldsymbol{\beta}(t)$  and  $\mathbf{Z}(t)$ . This model has been considered previously by several authors (Murhpy and Sen, 1991; Hastie and Tibshirani, 1990; Zucker and Lakatos, 1990; Cai and Sun, 2003; Winnett and Sasieni, 2003; Scheike and Martinussen, 2004). With covariates constant over time, the above model becomes the proportional hazards model (Cox, 1972) under the restriction that  $\boldsymbol{\beta}(t) = \boldsymbol{\beta}$ . When we take the risk sets to be fixed and known and conditional on a failure at time  $t$ , the probability that the failure concerns individual  $i$  is

$$\pi_i(\boldsymbol{\beta}(t), t) = Y_i(t) \exp\{\boldsymbol{\beta}(t)^T \mathbf{Z}_i(t)\} / \sum_{j=1}^n Y_j(t) \exp\{\boldsymbol{\beta}(t)^T \mathbf{Z}_j(t)\}, \quad i = 1, \dots, n. \quad (2)$$

The expectation and variance with respect to the probabilities  $\{\pi_i(\beta(t), t)\}_{i=1, \dots, n}$  are respectively a vector  $\mathbf{E}_{\beta(t)}(Z|t)$  of dimension  $p$  and a  $p \times p$  matrix  $\mathbf{V}_{\beta(t)}(Z|t)$  such that,

$$\begin{aligned}\mathbf{E}_{\beta(t)}(Z|t) &= \sum_{i=1}^n Z_i(t) \pi_i(\beta(t), t), \\ \mathbf{V}_{\beta(t)}(Z|t) &= \sum_{i=1}^n Z_i(t)^{\otimes 2} \pi_i(\beta(t), t) - \mathbf{E}_{\beta(t)}(Z|t)^{\otimes 2}.\end{aligned}$$

These quantities correspond to the conditional moments of the process  $\mathcal{Z}(t)$  for a fixed  $t$ , given the risk sets. The conditional variance-covariance matrix  $\mathbf{V}_{\beta(t)}(Z|t)$  is symmetric and positive definite. Thus, there exists an orthogonal matrix  $\mathbf{P}_{\beta(t)}(t)$  and a diagonal matrix  $\mathbf{D}_{\beta(t)}(t)$  such that

$$\mathbf{V}_{\beta(t)}(Z|t) = \mathbf{P}_{\beta(t)}(t) \mathbf{D}_{\beta(t)}(t) \mathbf{P}_{\beta(t)}(t)^T.$$

This leads us to define the symmetric matrix  $\mathbf{V}_{\beta(t)}(Z|t)^x$  by

$$\mathbf{V}_{\beta(t)}(Z|t)^x = \mathbf{P}_{\beta(t)}(t) (\mathbf{D}_{\beta(t)}(t))^x \mathbf{P}_{\beta(t)}(t)^T, \quad x \in \{-1/2, 1/2\}.$$

Denote

$$\mathbf{r}_{\beta(t)}(t) = \mathcal{Z}(t) - \mathbf{E}_{\beta(t)}(Z|t), \quad (3)$$

the residuals of the non-proportional hazards model (1) with parameter  $\beta(t)$  evaluated at time  $t$ . In the case of the proportional hazards model, these residuals reduce to the well-known Schoenfeld residuals (Schoenfeld 1982). Assume the case of a unique covariate ( $p = 1$ ) resulting in a univariate regression coefficient  $\beta(t)$ , a univariate conditional expectation  $E_{\beta(t)}(Z|t)$  and a univariate residual  $r_{\beta(t)}(t)$ . Consider the partial scores

$$U(\beta(t), t) = \int_0^t r_{\beta(s)}(s) d\bar{N}(s). \quad (4)$$

With a constant regression effect  $\beta$ , these correspond to the partial scores of Wei (1984). Wei was interested in goodness of fit for the two group problem and based a test on  $\sup_t |U(\hat{\beta}, t)|$ , large values indicating departures away from proportional hazards in the direction of non-proportional hazards. Considerable exploration of this idea, and substantial generalization via the use of martingale-based residuals, has been carried out by Lin et al. (1993, 1996) who showed that a wide choice of functions, potentially describing different kinds of departures from the model could be used. Apart from the two-group case, limiting distributions are complicated and usually approximated via simulation. Furthermore, Lin et al. (1993) pointed out that extensions of their methodology to the multivariate case or to the integration of time-dependent covariates are not straightforward. In order to overcome these difficulties, we follow the construction developed by Khmaladze (1981), working directly with the increments of the process rather than the process itself. We are then able to derive related processes for which the limiting

distributions are available analytically. To be more specific, when working with the ranks of the failure times and standardizing each increment of the process with a particular value rather than applying the same standardization for the whole score process, the limiting distribution of the multivariate process can be anticipated analytically and time-dependent covariates can be directly taken into account.

## 2.1 Time Scale

Let  $k_n = \#\{i : i = 1, \dots, n, \delta_i = 1, \det(\mathbf{V}_{\beta(X_i)}(Z|X_i)) > 0\}$ , where  $\#A$  denotes the cardinality of the set  $A$ , denote the number of observed failures such that the conditional variances assessed at the event-times are positive-definite matrices. In our setting, a null conditional variance at any time implies null conditional variances at later times. We assume that the number of failures  $k_n$  increases without bound as  $n$  increases without bound. By virtue of the fact that in Equation (1),  $\lambda_0(t)$  is unspecified, a monotonically increasing transformation of the times leaves inference for the regression parameter of the proportional hazards model unchanged. Therefore, Chauvel and O'Quigley (2014) considered the transformed times  $\phi_n(X_i)$  such that

$$\phi_n(X_i) = \frac{\bar{N}(X_i)}{k_n} \left( 1 + (1 - \delta_i) \frac{\#\{j : j = 1, \dots, n, X_j < X_i, \bar{N}(X_j) = \bar{N}(X_i)\}}{\#\{j : j = 1, \dots, n, \bar{N}(X_j) = \bar{N}(X_i)\}} \right). \quad (5)$$

Recall that the counting process  $\{\bar{N}(t)\}_{t \in \mathcal{T}}$  presents a unit jump at each observed failure time. On the new scale, the times in the set  $\{0, 1/k_n, 2/k_n, \dots, 1\}$  correspond to failure times, the  $i$ th ordered failure time, denoted  $t_i$ , is such that  $t_i = i/k_n$ . The set  $\{0, 1/k_n, 2/k_n, \dots, 1\}$  is included in but not equal to the set of images of all failure times. Censoring times can assume any value as long as they keep their original locations between adjacent failure times. For simplicity in Formula (5), we take these times to be spread uniformly between adjacent failure times, maintaining the original ranking. The time  $t_0$  on this scale corresponds to the  $100 \times t_0$ th percentile of failure in the sample. For instance, at time  $t_0 = 0.5$ , half of the failures are observed. The inverse transformation of  $\phi_n$  can be easily obtained and would enable us to interpret the results on the original time scale. On this transformed time scale, we can define the at-risk indicator  $Y_i^*(t)$  by  $Y_i^*(t) = I(\phi_n(X_i) \geq t)$  and the individual counting process  $N_i^*(t) = I(\phi_n(X_i) \leq t, \delta_i = 1)$ , for individual  $i = 1, \dots, n$ . In what follows, we only work with the standardized time scale, so that the process  $\mathcal{Z}$ , the expectation  $\mathbf{E}_{\beta(t)}(Z|t)$  and the variance  $\mathbf{V}_{\beta(t)}(Z|t)$ , of which extensions are straightforward, are defined for  $0 \leq t \leq 1$ . Define the counting process associated with the transformed times which has unit jumps at failure-times on the new scale by

$$\bar{N}^*(t) = \sum_{i=1}^n I(\phi_n(X_i) \leq t, \delta_i = 1), \quad 0 \leq t \leq 1.$$

On the new time scale, the partial scores (4) can be re-expressed as

$$\mathbf{U}(\beta(t), t) = \int_0^t \mathbf{r}_{\beta(s)}(s) d\bar{N}^*(s) = \sum_{i=1}^{\lfloor k_n t \rfloor} \mathbf{r}_{\beta(t_i)}(t_i), \quad 0 \leq t \leq 1,$$

where the  $i$ th element of the vector  $\int_0^t \mathbf{a}(s) d\bar{N}^*(s)$  is  $\int_0^t \mathbf{a}_i(s) d\bar{N}^*(s)$  for any  $\mathbb{R}^p$ -valued  $\mathbf{a}(t) = (a_1(t), \dots, a_p(t))$ ,  $i = 1, \dots, p$  and  $\lfloor x \rfloor$  gives the largest integer less than or equal to  $x$ .

## 2.2 Multivariate Standardized Score Process

Before defining the standardized score process, let us give the assumptions needed in the sequel. Let  $t \in [0, 1]$ ,  $\gamma(t)$  be a regression function, not necessarily equals to  $\beta(t)$  and

$$\begin{aligned} S^{(0)}(\gamma(t), t) &= n^{-1} \sum_{i=1}^n Y_i(t) e^{\gamma(t) Z_i(t)}, & \mathbf{S}^{(1)}(\gamma(t), t) &= n^{-1} \sum_{i=1}^n Y_i(t) Z_i(t) e^{\gamma(t) Z_i(t)}, \\ \mathbf{S}^{(2)}(\gamma(t), t) &= n^{-1} \sum_{i=1}^n Y_i(t) Z_i(t)^{\otimes 2} e^{\gamma(t) Z_i(t)}. \end{aligned}$$

Using these notations, we have the equalities  $\mathbf{E}_{\gamma(t)}(Z|t) = \mathbf{S}^{(1)}(\gamma(t), t)/S^{(0)}(\gamma(t), t)$  and  $\mathbf{V}_{\gamma(t)}(Z|t) = \mathbf{S}^{(2)}(\gamma(t), t)/S^{(0)}(\gamma(t), t) - \mathbf{E}_{\gamma(t)}(Z|t)^{\otimes 2}$ . Notice that the Jacobian matrix of  $\mathbf{E}_{\gamma(t)}(Z|t)$  is the variance-covariance matrix  $\mathbf{V}_{\gamma(t)}(Z|t)$ . Consider that the following assumptions, similar to those of Andersen and Gill (1982) hold:

- A. (Asymptotic stability). There exists a neighbourhood  $\mathcal{B}$  of  $\beta(t)$  and vector and matrix functions  $\mathbf{s}^{(r)}(\gamma(t), t)$ ,  $r = 0, 1, 2$ , defined for  $t \in [0, 1]$  and  $\gamma(t) \in \mathcal{B}$  such that  $\mathbf{0}$  and  $\beta(t)$  are in the interior of  $\mathcal{B}$ , for all  $t \in [0, 1]$  and

$$\sqrt{n} \sup_{t \in [0, 1], \gamma(t) \in \mathcal{B}} \|\mathbf{S}^{(r)}(\gamma(t), t) - \mathbf{s}^{(r)}(\gamma(t), t)\| \xrightarrow[n \rightarrow \infty]{\mathbb{P}} 0.$$

- B. (Asymptotic regularity). All functions defined in assumption A. are uniformly continuous in  $t \in [0, 1]$ . In addition, for  $r = 0, 1, 2$ ,  $\mathbf{s}^{(r)}(\gamma(t), t)$  are continuous functions of  $\gamma(t) \in \mathcal{B}$ , bounded on  $\mathcal{B} \times [0, 1]$  and  $\mathbf{s}^{(0)}(\gamma(t), t)$  is bounded away from 0.

- C. (Homoscedasticity). There exists a symmetric and positive definite matrix  $\Sigma$  and a series of positive constants  $(M_n)_n$  converging to 0 as  $n$  goes to infinity such that

$$\begin{aligned} \sup_{t \in [0, 1], \gamma(t) \in \mathcal{B}} \left\| \frac{\partial}{\partial \beta} \mathbf{V}_{\beta}(Z|t) \Big|_{\beta=\gamma(t)} \right\| &\leq M_n \quad a.s., \\ \sup_{t \in [0, 1], \gamma(t) \in \mathcal{B}} \|\mathbf{V}_{\gamma(t)}(Z|t) - \Sigma\| &\xrightarrow[n \rightarrow \infty]{\mathbf{L}^1} 0. \end{aligned}$$



By analogy with the empirical quantities, we denote  $e(\gamma(t), t) = s^{(1)}(\gamma(t), t)/s^{(0)}(\mathbf{0}, t)$  and  $v(\gamma(t), t) = s^{(2)}(\gamma(t), t)/s^{(0)}(\mathbf{0}, t) - e(\gamma(t), t)^{\otimes 2}$ .

The two first conditions are classical and introduced by Andersen and Gill (1982) for using counting process and martingales theory, such as Lengart's inequality or Rebolledo's theorem. Although we use a different approach here, that we believe is simpler to comprehend, the same assumptions are made. Notice that  $\mathbf{V}_{\gamma(t)}(Z|t)$  is, by definition, the sample-based variance of  $Z$  given  $T = t$  under the model with parameter  $\gamma(t)$ . Thus, condition C. of homoscedasticity means that the asymptotic variance does not depend on time. This condition is implicitly used in the context of the proportional hazards regression, for instance when estimating the variance of the parameter  $\beta$  or when applying the log-rank test. The contribution to the global variance is the same at each failure time, by the use of an unweighted sum of each term. This stability of variance has also been noticed by several authors, for example Grambsch and Therneau (1994). From the previous section, under the non-proportional hazards model (1), the increments of the process  $\sum_{j=1}^{\lfloor k_n t \rfloor} \mathcal{Z}(t_j)$  at  $t = t_i$  have mean  $\mathbf{E}_{\beta(t_i)}(Z|t_i)$  and variance-covariance matrix  $\mathbf{V}_{\beta(t_i)}(Z|t_i)$ . The increments of the process are independent, either by design in view of the conditional model, or by the arguments of Cox (1975). Thus only the existence of the variance is necessary to carry out appropriate standardization and to be able to appeal to the functional central limit theorem. This leads us to define a standardized version of the multivariate score process:

**Definition 1** *The multivariate standardized score process evaluated at parameter  $\beta_0$  and at failure time  $t \in \{0, 1/k, 2/k_n, \dots, 1\}$  is*

$$\mathbf{U}^*(\beta_0, t) = \frac{1}{\sqrt{k_n}} \int_0^t \mathbf{V}_{\beta_0}(Z|s)^{-1/2} \mathbf{r}_{\beta_0}(s) d\bar{N}^*(s) = \frac{1}{\sqrt{k_n}} \sum_{i=1}^j \mathbf{V}_{\beta_0}(Z|t_i)^{-1/2} \mathbf{r}_{\beta_0}(t_i).$$

The  $\mathbf{U}^*$  process is only defined on  $k_n$  equispaced points of the interval  $[0, 1]$  but we extend our definition to the whole interval via linear interpolation so that, for  $u$  in the interval  $[t_j, t_{j+1}]$ , we write

$$\mathbf{U}^*(\beta_0, u) = \mathbf{U}^*(\beta_0, t_j) + \{uk_n - j\} \{\mathbf{U}^*(\beta_0, t_{j+1}) - \mathbf{U}^*(\beta_0, t_j)\}.$$

The following theorem gives the asymptotic behaviour of  $\mathbf{U}^*(\beta_0, \cdot)$ :

**Theorem 1** *Under the non-proportional hazards model of parameter  $\beta(t)$ , we have the following convergence in distribution:*

$$\mathbf{U}^*(\beta_0, \cdot) - \sqrt{k_n} \mathbf{C}_n \xrightarrow[n \rightarrow \infty]{\mathcal{D}} \mathbf{W}_p, \quad (6)$$

where  $\mathbf{W}_p$  is a standard Brownian motion of dimension  $p$  and, for all  $t \in [0, 1]$ ,

$$\mathbf{C}_n(t) = \frac{1}{k_n} \sum_{i=1}^{\lfloor tk_n \rfloor} \mathbf{V}_{\beta_0}(Z|t_i)^{-1/2} \{\mathbf{E}_{\beta_0}(Z|t_i) - \mathbf{E}_{\beta(t_i)}(Z|t_i)\}.$$

In addition, we have the convergence of probability

$$\sup_{t \in [0,1]} \left\| \mathbf{C}_n(t) - \boldsymbol{\Sigma}^{1/2} \int_0^t \{\boldsymbol{\beta}(s) - \boldsymbol{\beta}_0\} ds \right\| \xrightarrow[n \rightarrow \infty]{P} 0, \quad (7)$$

where  $\int_0^t a(s) ds = \left( \int_0^t a_1(s) ds, \dots, \int_0^t a_p(s) ds \right)$  for any  $\mathbb{R}^p$ -valued function  $a = (a_1, \dots, a_p)$ .

The proof is given in Appendix A and is based on the multivariate functional central limit theorem of Helland (1982). The second term of formula (6) increases without bound as the sample size goes to infinity. In practical situations, when the model generating the observations is based on  $\boldsymbol{\beta}(t)$ , Theorem 1 in addition to Slutsky's lemma indicate that  $\mathbf{U}^*(\boldsymbol{\beta}_0, \cdot)$  will look like a multivariate Brownian motion with an added drift term:

**Corollary 1** *Under the model (1) with parameter  $\boldsymbol{\beta}(t)$ , we have, for all  $\boldsymbol{\beta}_0$ ,*

$$\mathbf{U}^*(\boldsymbol{\beta}_0, \cdot) - \sqrt{k_n} \boldsymbol{\Sigma}^{1/2} IB \xrightarrow[n \rightarrow \infty]{\mathcal{D}} \mathbf{W}_p,$$

where  $IB(t) = \int_0^t \{\boldsymbol{\beta}(s) - \boldsymbol{\beta}_0\} ds$ ,  $0 \leq t \leq 1$ . In addition,  $\hat{\boldsymbol{\Sigma}} = k_n^{-1} \sum_{i=1}^{k_n} \mathbf{V}_{\boldsymbol{\beta}_0}(Z|t_i)$  is a consistent estimator of  $\boldsymbol{\Sigma}$ , and

$$\hat{\boldsymbol{\Sigma}}^{-1/2} \mathbf{U}^*(\boldsymbol{\beta}_0, \cdot) - \sqrt{k_n} IB \xrightarrow[n \rightarrow \infty]{\mathcal{D}} \boldsymbol{\Sigma}^{-1/2} \mathbf{W}_p. \quad (8)$$

In the sequel, the standardized score process is evaluated in  $\boldsymbol{\beta}_0 = \mathbf{0}$ , where  $\mathbf{0}$  is the null vector of  $\mathbb{R}^p$ . As a consequence, the plot of  $\hat{\boldsymbol{\Sigma}}^{-1/2} \mathbf{U}^*(\mathbf{0}, t)$  against the time  $t$  gives an indication on the shape of  $\int_0^t \boldsymbol{\beta}(s) ds$  which is reflected by the shape of the drift of the process (equation (8)). In the univariate case ( $p = 1$ ) or in the multivariate case with independent covariates, the process  $\mathbf{U}^*(\mathbf{0}, \cdot)$  can be directly plotted over time, with no additional standardization since  $\boldsymbol{\Sigma}$  is the identity matrix. However, when dealing with correlated covariates, a global standardization is needed for isolating each effect  $\beta_i(t)$  on each process  $\left[ \hat{\boldsymbol{\Sigma}}^{-1/2} \mathbf{U}^*(\boldsymbol{\beta}_0, \cdot) \right]_i$ ,  $i = 1, \dots, p$ . A linear drift corresponds to a constant over time regression effect. Our proposed method takes into account the covariances between all covariates and the goodness-of-fit of the overall model is directly evaluated instead of checking proportionality of hazards for one covariate at a time.

Illustrations are given in the univariate case. Figure 1(a) represents a simulation of the process  $U^*(0, t)$  over time  $t$ , under the model with a null regression parameter  $\beta(t) = 0$ . Even under moderate to small sample size, the Brownian motion approximation appears accurate enough for reliable inference. Consider a proportional hazards model with  $\beta(t)$  constant over time but not null. Corollary 1 suggests that a good approximation for this process is a Brownian motion with a linear drift. An indication of the plausibility of this is shown in Figure 1(b), where  $\beta(t)$  is set to 0.5. Departures from the proportional hazards assumption can be of various forms. For instance, the effect can be constant and then decreasing after some time  $\tau$ , the effect can be piecewise constant

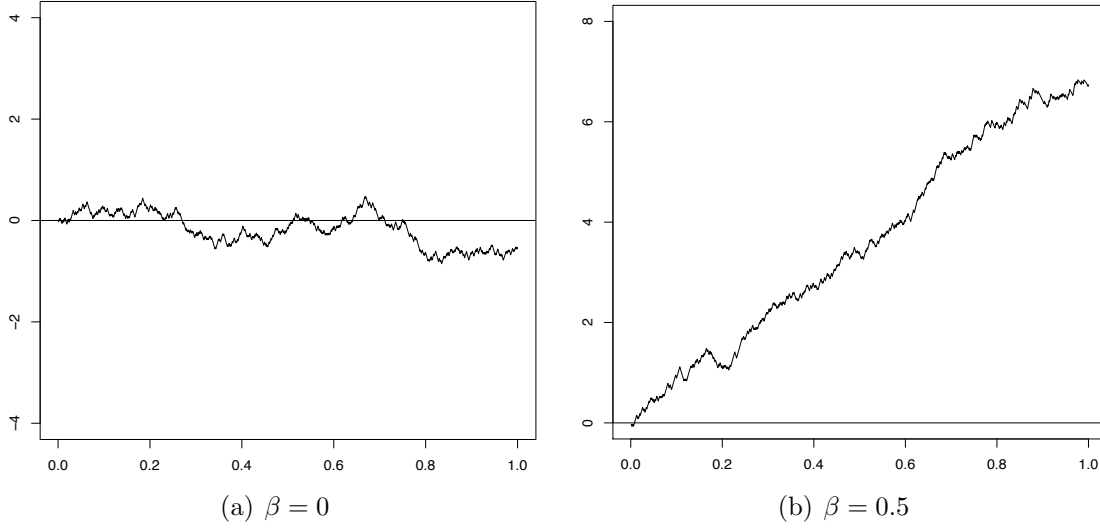


Figure 1: Processes  $U^*(0, t)$  based on data simulated from proportional hazards models of parameter  $\beta$ .

over time or it can increase over time. Corollary 1 implies that the shape of the drift of the process  $U^*(0, \cdot)$  will reflect the shape of the cumulated regression coefficient. As an illustration, Figure 2(a) represents a simulated process under the non-proportional hazards model, with  $\beta(t)$  piecewise constant. Before  $t = 0.5$ , there is a linear trend corresponding to  $\beta = 1$  and for  $t > 0.5$ ,  $\beta$  equals zero and the process  $U^*(0, t)$  is constant in expectation over time. Figure 2(b) represents a simulated standardized score process for a changepoint model with the regression parameter  $\beta(t) = I(t \leq 1/3) + 0.5I(t > 2/3)$ . The trend of the process can be separated into 3 straight lines reflecting the strength of the effect: the slope of the first part seems twice higher than the one of the last part and the slope of the second part is null. The following proposition enables the construction of a confidence band for each process:

**Proposition 1** *Let  $i = 1, \dots, p$ . Consider the hypothesis  $H_{0,i} : \exists b_i, \beta_i(t) = b_i$  and its alternative  $H_{1,i} : \nexists b_i, \beta_i(t) = b_i$ . Under the model (1) of parameter  $\beta(t) = (\beta_1(t), \dots, \beta_p(t))$  not necessarily equals to  $\beta_0$  and  $H_{0,i}$ , we have, for all  $a \geq 0$ ,*

$$\begin{aligned} & \lim_{n \rightarrow +\infty} P \left( \left\| \hat{\Sigma}_{\cdot, i}^{-1/2} \right\|_2^{-1} \sup_{t \in [0, 1]} \left| \left( \hat{\Sigma}^{-1/2} \{ \mathbf{U}^*(\beta_0, t) - t \mathbf{U}^*(\beta_0, 1) \} \right)_i \right| \leq a \right) \\ &= P \left( \sup_{t \in [0, 1]} |B(t)| \leq a \right), \end{aligned} \quad (9)$$

where  $B$  is a Brownian bridge and  $\left\| \hat{\Sigma}_{\cdot, i}^{-1/2} \right\|_2 = \left( \sum_{j=1}^p \left( \hat{\Sigma}_{j, i}^{-1/2} \right)^2 \right)^{1/2}$ . Therefore, by

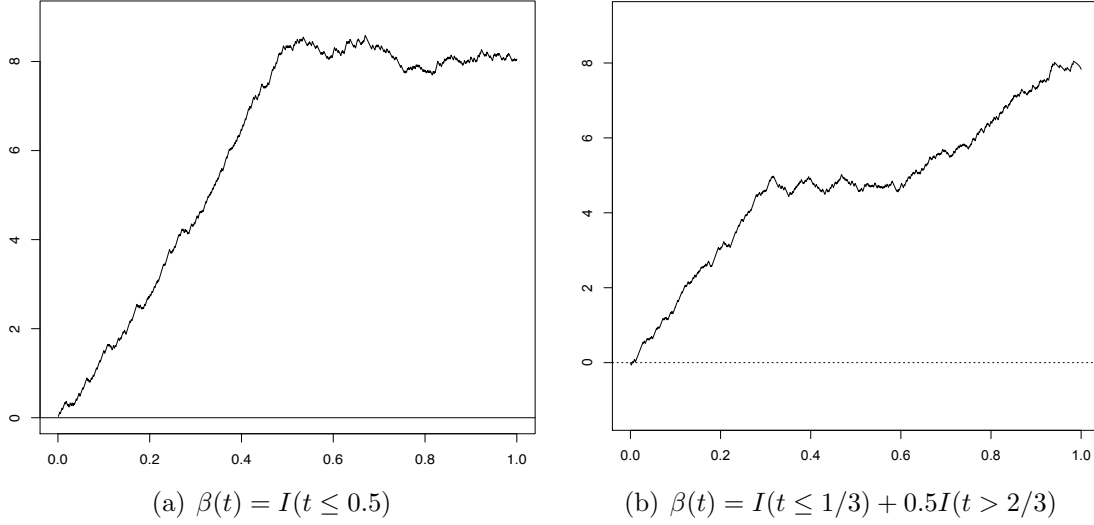


Figure 2: Processes  $U^*(0, t)$  based on data simulated from models (1) of parameter  $\beta(t)$ .

denoting  $a(\alpha)$  the quantile of order  $\alpha$  of the Kolmogorov distribution, we have

$$\lim_{n \rightarrow +\infty} P\left(\forall t \in [0, 1], \left[\hat{\Sigma}^{-1/2} \mathbf{U}^*(\beta_0, t)\right]_i \in IC_i(\alpha)\right) = 1 - \alpha,$$

with

$$IC_i(\alpha) = \left[t \left[\hat{\Sigma}^{-1/2} \mathbf{U}^*(\beta_0, 1)\right]_i - \left\|\hat{\Sigma}_{\cdot, i}^{-1/2}\right\|_2 a(\alpha); t \left[\hat{\Sigma}^{-1/2} \mathbf{U}^*(\beta_0, 1)\right]_i - \left\|\hat{\Sigma}_{\cdot, i}^{-1/2}\right\|_2 a(\alpha)\right].$$

The proof can be found in Appendix B. If the  $i$ th element of the process  $\hat{\Sigma}^{-1/2} \mathbf{U}^*(\beta_0, t)$  leaves the confidence band  $IC(\alpha)_i$ , we reject the hypothesis that the effect  $\beta_i(t)$  is constant over time with an asymptotic level of  $\alpha$ . However, when testing simultaneously several hypotheses  $H_{0,i}$  of constant effects for different covariates, the global type I error is inflated. This means that one process could leave its confidence band whereas the corresponding effect is constant over time with a level higher than  $\alpha$ . This does not seem to be a problem since the plot of the confidence interval is just one of the tools we use to select the variables respecting the proportional hazards assumption. We do not base a definitive conclusion regarding this assumption on this confidence band only, and the non-detection of a constant effect will be corrected with the other steps of the selection variable method we propose in this article. Of course, corrections for multiple testings could be applied.

Whether effects are of a proportional hazards or a non-proportional hazards form, essentially, all of the information concerning the regression effect  $\beta(t)$  is captured in the process  $\mathbf{U}^*(\mathbf{0}, \cdot)$ . The process allows the data to speak for themselves, not unlike a scatterplot in linear regression, in which trends and non-linearity may be apparent, since we evaluate the process at  $\beta_0 = \mathbf{0}$ . No parameter has to be estimated and expectations

and variance-covariance matrices are the usual sequential empirical quantities. The process, based on the residuals of the non-proportional hazards model, is a useful tool in the evaluation of its goodness-of-fit. These residuals can also be used in the construction of a predictive accuracy measure of the model.

### 3 INTERPLAY OF FIT AND PREDICTION

#### 3.1 $R^2$ Coefficient as a Measure of Predictive Ability

For any random variables  $X$  and  $Y$  having second moments, the formula

$$\text{Var}(Y) = E(\text{Var}(Y|X)) + \text{Var}(E(Y|X)), \quad (10)$$

leads to the natural definition of explained variation as the ratio of the variance of the expected values of the response variable under the model given the explanatory variables to the marginal variance of the response variable. In light of the Chebyshev inequality, we see that explained variation directly quantifies predictive strength.

In the non-proportional hazards model (1) with one covariate  $Z(t)$  ( $p = 1$ ), the explained variation makes use of the variance decomposition given in equation (10) in which  $Y$  is replaced by  $Z(t)$  and  $X$  by  $T$ , leading to the definition:

**Definition 2** *In the univariate non-proportional hazards model (1), the explained variation, expressed as a function of the time-dependent regression coefficient  $\beta(t)$ , is defined by*

$$\Omega^2(\beta(t)) = \frac{\text{Var}(E(Z|T))}{\text{Var}(Z)} = 1 - \frac{E(\text{Var}(Z|T))}{\text{Var}(Z)}.$$

In the multivariate non-proportional hazards model (1), individual  $i$  is characterized by its real-valued prognostic index  $\eta_i(t) = \beta(t)^T \mathbf{Z}_i(t)$ , being a realization of  $\eta(t) = \beta(t)^T \mathbf{Z}(t)$ . Therefore it is equivalent to evaluate the quality of prediction of the model via  $\mathbf{Z}$  or  $\eta$ . We adopt the latter possibility.

**Definition 3** *The explained variation of the non-proportional hazards model (1) with multiple covariates can be defined by a function of the time-dependent regression coefficient  $\beta(t)$  by*

$$\Omega^2(\eta(t)) = \frac{\text{Var}(E(\eta|T))}{\text{Var}(\eta)} = 1 - \frac{E(\text{Var}(\eta|T))}{\text{Var}(\eta)}, \quad \eta(t) = \beta(t)^T \mathbf{Z}(t).$$

Some properties of  $\Omega^2$  can be found in O'Quigley (2008, chap. 13) or in O'Quigley and Xu (2012, chap. 27). In these book chapters,  $\Omega^2$  is a function of a constant regression parameter  $\beta$ , corresponding to the proportional hazards model. Extension to a time-dependent regression parameter is straightforward. The explained variation coefficient remains constant when applying a monotonically increasing transformation on

time. Thus, we work on the standardized time scale as described in Section 2.2.

The explained variation is a population parameter that needs to be estimated. Several estimators have been proposed in the literature (Choodari-Oskooei et al. 2012). Our goal here is not to present an exhaustive review of these estimators. We focus on the  $R^2$  coefficient introduced by O'Quigley and Flandre (1994) since it is built with the same residuals as the standardized score process. We recall its definition by extending it to the non-proportional hazards case. Let us define the expectation over time of the expected squared discrepancy between the covariate or prognostic index evaluated with parameter  $\alpha_2(t)$  and their expected value under the non-proportional hazards model (1) of parameter  $\alpha_1(t)$ , not necessarily equals to  $\alpha_2(t)$ , of dimension  $p$

$$Q(F, \alpha_1(t), \alpha_2(t)) = \begin{cases} \int_0^1 E_{\alpha_1(t)} \left( \left\{ \alpha_2(t)^T Z(t) - E_{\alpha_1(t)} (\alpha_2(t)^T Z \mid T = t) \right\}^2 \middle| T = t \right) dF(t) & \text{if } p > 1 \\ \int_0^1 E_{\alpha_1(t)} \left( \left\{ Z(t) - E_{\alpha_1(t)} (Z \mid T = t) \right\}^2 \middle| T = t \right) dF(t) & \text{if } p = 1, \end{cases}$$

where  $F$  is the cumulative distribution function of  $T$ . Then  $\Omega^2(\beta(t))$  can be expressed as

$$\Omega^2(\beta(t)) = 1 - \frac{Q(F, \beta(t), \beta(t))}{Q(F, \mathbf{0}, \beta(t))}. \quad (11)$$

Let us denote  $\hat{F}$  the estimator of the cumulative distribution function of  $T$  such that  $\hat{F}(t) = k_n^{-1} \bar{N}^*(t)$ .  $\hat{F}$  corresponds to the usual empirical cumulative distribution function of  $T$  in the uncensored case. Then,  $Q(F, \alpha_1(t), \alpha_2(t))$  can be estimated by

$$\hat{Q}(\hat{F}, \alpha_1(t), \alpha_2(t)) = \begin{cases} \int_0^1 \{ \alpha_2(s)^T r_{\alpha_1(s)}(s) \}^2 d\hat{F}(s) = \frac{1}{k_n} \sum_{i=1}^{k_n} \{ \alpha_2(t_i)^T r_{\alpha_1(t_i)}(t_i) \}^2 & \text{if } p > 1 \\ \int_0^1 \{ r_{\alpha_1(s)}(s) \}^2 d\hat{F}(s) = \frac{1}{k_n} \sum_{i=1}^{k_n} \{ r_{\alpha_1(t_i)}(t_i) \}^2 & \text{if } p = 1. \end{cases}$$

The  $R^2$  coefficient can then be defined by  $R^2 = R^2(\hat{\beta}(t))$ , where, for all vector  $\alpha(t)$

of dimension  $p$ ,

$$R^2(\boldsymbol{\alpha}(t)) = 1 - \frac{\hat{Q}(\hat{F}, \boldsymbol{\alpha}(t), \boldsymbol{\alpha}(t))}{\hat{Q}(\hat{F}, \mathbf{0}, \boldsymbol{\alpha}(t))} = \begin{cases} 1 - \frac{\sum_{i=1}^{k_n} \{\boldsymbol{\alpha}(t_i)^T r_{\boldsymbol{\alpha}(t_i)}(t_i)\}^2}{\sum_{i=1}^{k_n} \{\boldsymbol{\alpha}(t_i)^T r_{\mathbf{0}}(t_i)\}^2} & \text{if } p > 1 \\ 1 - \frac{\sum_{i=1}^{k_n} r_{\boldsymbol{\alpha}(t_i)}(t_i)^2}{\sum_{i=1}^{k_n} r_{\mathbf{0}}(t_i)^2} & \text{if } p = 1. \end{cases} \quad (12)$$

The explained variation coefficient  $\Omega^2(\boldsymbol{\beta}(t))$  can be estimated by  $R^2(\hat{\boldsymbol{\beta}}(t))$ , where  $\hat{\boldsymbol{\beta}}(t)$  is a consistent estimator of the true value of the regression coefficient  $\boldsymbol{\beta}(t)$ . The following theorem will be useful to evaluate the goodness of fit of the model (1).

**Theorem 2** *Under the non-proportional hazards model (1) of parameter  $\boldsymbol{\beta}(t)$ , we have the following convergence*

$$|R^2(\boldsymbol{\beta}(t)) - R^2(\hat{\boldsymbol{\beta}}(t))| \xrightarrow[n \rightarrow +\infty]{a.s.} 0,$$

and, if  $p = 1$ , with probability one,

$$\arg \max_{b(t)} \lim_{n \rightarrow +\infty} R^2(b(t)) = \boldsymbol{\beta}(t).$$

The proof is given in appendix C. The theorem states that if  $\boldsymbol{\beta}(t)$  is the true regression coefficient, the maximum of  $R^2$  is well approximated by  $R^2(\hat{\boldsymbol{\beta}}(t))$  for a large enough sample size. This result has been shown for one covariate in the model, and in the case of multiple covariates, we conjecture an analogous result for the one-dimensional prognostic index. The predictive ability measure and the standardized score process are built with the same ingredients. The standardized score process enables to check the fit of the model, whereas the  $R^2$  coefficient is a measure of the predictive ability of the model. Although different, these two aspects of the model are related; their construction with the same quantities seems then to be quite natural.

### 3.2 Using the $R^2$ Coefficient to Improve the Fit

Using the results of Theorem 1 and its corollary, the standardized score process can be used to determine the shape of the temporal regression effect. No other tools such as smoothing, the projection on a basis of functions or kernel estimation are needed (Cai and Sun, 2003; Hastie and Tibshirani, 1990; Scheike and Martinussen, 2004). For instance, as shown in Figure 2(a), a constant effect until time  $\tau$  followed by a null effect is easily detectable, especially with moderate and larger sample sizes. Assume that the time-dependent regression parameter can be expressed as  $\boldsymbol{\beta}(t) = (\beta_1(t), \dots, \beta_p(t))$ , where

$\beta_j(t) = \beta_{0,j} B_j(t)$  ( $j = 1, \dots, p$ ) with  $\beta_0 = (\beta_{0,1}, \dots, \beta_{0,p}) \in \mathbb{R}^p$  an unknown regression parameter and  $\mathbf{B} = (B_1, \dots, B_p)$  a known  $\mathbb{R}^p$ -valued function of the time. Thus,  $\widehat{\beta}(t) = (\widehat{\beta}_1(t), \dots, \widehat{\beta}_p(t))$  where  $\widehat{\beta}_j(t) = \widehat{\beta}_{0,j} B_j(t)$  ( $j = 1, \dots, p$ ) with  $\widehat{\beta}_0$  the maximum likelihood estimator of  $\beta_0$ , obtained via classical maximization of the partial likelihood (Cox, 1972). The function  $B(t)$  can be determined graphically using the standardized score process (see the examples of Section 4). In addition, the confidence bands defined in Proposition 1 can help to evaluate the plausibility of a constant effect  $\beta_j(t)$  over time resulting in a constant function  $B_j$ , for each  $j = 1, \dots, p$ .

When dealing with non-proportional hazards, the investigator needs an instrument other than one that is focused solely on fit. This can be provided by the  $R^2$  coefficient that not only indicates predictive strength but will tend to a maximum value when the correct form of  $B(t)$  is chosen (Theorem 2). When different competing models provide plausible forms for  $B(t)$ , the one maximizing the  $R^2$  coefficient would be considered the best. Using this procedure, we obtain a non-proportional hazards model with a good fit and a maximal predictive ability. The predictive ability measure is maximized on the set  $\mathcal{B}$  of the temporal regression effects selected by the investigator. Formally, let  $\mathcal{B} = \{\beta_1(t), \dots, \beta_m(t)\}$  be a set of  $m$  functions from  $[0, 1]$  to  $\mathbb{R}^p$ . The selected regression function  $\beta^*(t)$  is such that

$$\beta^*(t) = \arg \max_{b(t) \in \mathcal{B}} R^2(b(t)).$$

The following theorem gives an equivalence between this maximization problem when  $n \rightarrow \infty$  and a problem of minimization of  $L^2$  norms.

**Theorem 3** *Let  $p = 1$ . Under the non-proportional hazards model (1) with regression parameter  $\beta(t)$  not necessarily in  $\mathcal{B}$ , asymptotically,  $\beta^*(t) = \arg \max_{\alpha(t) \in \mathcal{B}} \lim_{n \rightarrow \infty} R^2(\alpha(t))$  is the solution of*

$$\beta^*(t) = \arg \min_{\alpha(t) \in \mathcal{B}} \|\beta(t) - \alpha(t)\|_{2,W},$$

where  $\|a(t)\|_{2,W} = \left( \int_0^1 a(t)^2 v(c(t), t)^2 dt \right)^{1/2}$  is a weighted  $L^2$  norm of the function  $a(t)$  from  $[0, 1]$  to  $\mathbb{R}$ , with  $c(t)$  lying between 0 and  $a(t)$ .

Proof can be found in Appendix D. In other words, for large enough sample sizes, selecting the regression coefficient by maximizing the  $R^2$  coefficient is the same as selecting the closest temporal regression function to the true coefficient in the  $L^2$  norm sense.

A model is chosen to fit a dataset because of either a good fit or a good predictive capacity. Several models could present one of these aspects or both of them, not only the "true" model. Priority is given to the goodness of fit, with the selection of possible time-dependent coefficients, and in a second phase, the predictive capacity is considered.



We have chosen to work with the  $R^2$  coefficient but notice that other predictive ability measures verifying Theorem 2 might be considered. When the trend of the process is a concave function, the effect diminishes over time, whereas in presence of a convex function, the effect increases. In order to obtain the largest possible  $R^2$ , we could create a temporal effect matching more and more closely the observed trend of the process, e.g. piecewise constant effects with multiple changepoints. In general, this would result in an overfit. In this case, the interpretation of the coefficient is not clear. A tradeoff has to be established between a high predictive ability and the simplicity of the coefficient, especially regarding its interpretation. This parallels linear regression where the estimated explained variation is positively biased and this bias increases with the dimension of the model. Some balance needs to be struck between the goal of improved prediction and the dangers of over optimistic predictions as a result of over fitting.

## 4 SOME SIMULATED EXAMPLES

The simulations are performed with a moderate sample size set to  $n = 200$  subjects and  $\lambda_0(t) = 1$ . All cases presented here are uncensored. The effect of an independent censoring mechanism on the process is the same as a reduction in the sample size.

### 4.1 Univariate cases

In both considered cases, the covariate follows a Bernoulli distribution of parameter 0.5. First, we consider the proportional hazards situation by setting  $\beta(t) = 1.5$ . The

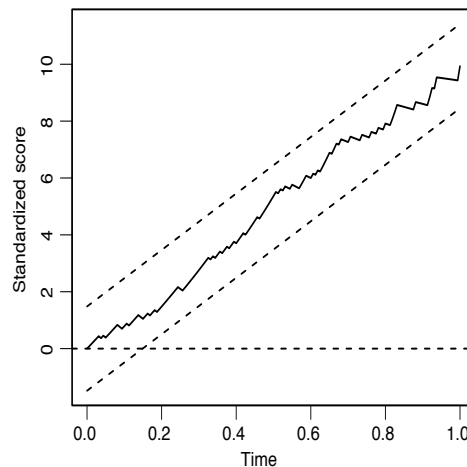


Figure 3: Standardized score process  $U^*(0, \cdot)$  (solid line) and confidence bands (dotted lines) on a simulated dataset with constant regression coefficient  $\beta(t) = 1.5$ .

standardized score process  $U^*(0, \cdot)$  (solid line) and its confidence bands under propor-

tional hazards assumption (dotted lines) are plotted over time in Figure 3. A drift is observed, the effect is not null. The drift seems linear and the process stays between the confidence bands: the hypothesis of a proportional hazards model seems reasonable. The usual maximum partial likelihood estimator is estimated at 1.59 which gives an  $R^2$  of 0.35.

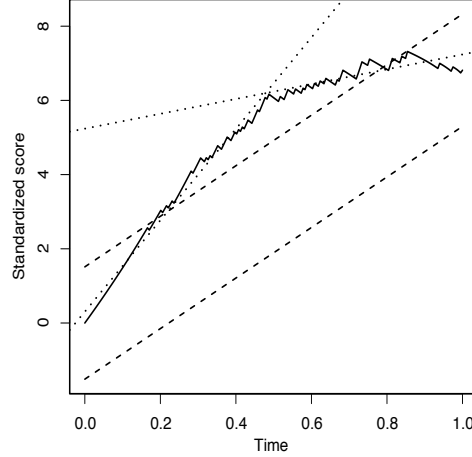


Figure 4: Standardized score process  $U^*(0, \cdot)$  (solid line), confidence bands (dashed lines) and a fitted changepoint model (dotted lines) on a simulated dataset with  $\beta(t) = 3(1 - t)^2$ .

The next case deals with a smooth decreasing effect. We simulate a dataset with  $\beta(t) = 3(1 - t)^2$ . The resulting standardized score process  $U^*(0, \cdot)$  (solid line) is plotted over time in Figure 4 with its confidence bands under proportional hazards assumption (dotted lines). The process leaves the confidence bands which indicates that the proportional hazards assumption does not hold. The concavity of the trend gives an indication regarding the decrease of the effect. Amongst other possibilities, the effect could be linear, of a quadratic shape or a piecewise constant function of the time. In the latter case, the trend appears linear up to time  $t = 0.5$  corresponding to a constant coefficient. Then, the drift changes to a lower constant value, corresponding to a coefficient  $\tilde{\beta}(t) = \beta_0\{I(t \leq 0.5) + C.I(t > 0.5)\}$ , where  $\beta_0$  and  $C$  are unknown.  $C$  is the value by which the coefficient is multiplied in the second part of the study. In Figure 4, using linear regression, two straight dotted lines have been fitted to the process, before and after the changepoint time  $t = 0.5$ . The ratio of the second slope over the first one is the value  $C = 0.16$ . Various models with decreasing effect  $\beta(t) = \beta_0 B(t)$  have been selected, their  $R^2$  coefficients and  $\hat{\beta}_0$  the maximum partial likelihood estimator of  $\beta_0$  have been evaluated in Table 1. The lowest  $R^2$  coefficient corresponds to the proportional hazards model and the largest  $R^2$  coefficient of Table 1 is the one associated with the model of regression coefficient  $\beta(t) = \beta_0(1 - t)^2$ , with an estimation of  $\beta_0$  equals to 0.37. Using our procedure, the regression coefficient used to create the dataset has been selected.

$\beta(t)$	$\beta_0$	$\beta_0(1-t)$	$\beta_0(1-t)^2$	$\beta_0(1-t^2)$	$\tilde{\beta}(t)$
$\hat{\beta}_0$	1.06	2.45	3.73	1.77	1.83
$R^2$	0.25	0.36	0.37	0.34	0.34

Table 1: Maximum partial likelihood estimators  $\hat{\beta}(t)$  and  $R^2$  coefficients on a simulated dataset with  $\beta(t) = 3(1-t)^2$ .

## 4.2 Multivariate case

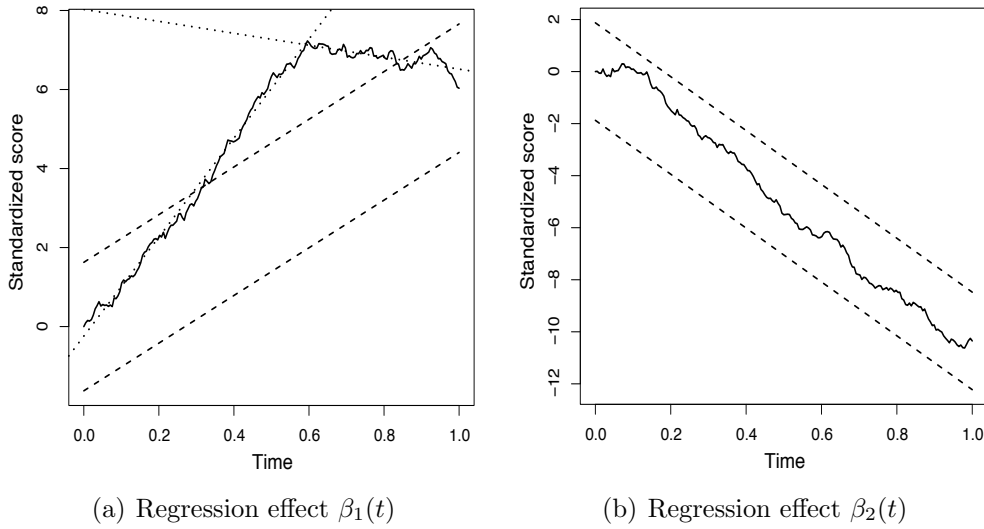


Figure 5: Standardized score process  $\hat{\Sigma}^{-1}\mathbf{U}^*(0, \cdot)$  (solid line), confidence bands (dashed lines) and a fitted changepoint model (dotted lines) on a simulated dataset with  $\beta_1(t) = I(t \leq 0.5)$  and  $\beta_2(t) = -1$ .

We simulate two standard normal covariates  $Z^1$  and  $Z^2$  with covariance equals to 0.5. We set  $\beta_1(t) = I(t \leq 0.5)$  and  $\beta_2(t) = 1$ . Each component of the bivariate process  $\hat{\Sigma}^{-1/2}\mathbf{U}^*(\mathbf{0}, \cdot)$  (solid lines) is plotted over time on Figures 5(a) and 5(b) with the confidence bands (dotted lines). Clearly, the proportional hazards assumption is rejected for covariate  $Z^1$  since the process leaves the confidence band. The shape of the process indicates a piecewise constant regression coefficient, with a changepoint at time  $t = 0.6$ . As in the univariate case, two straight (dashed) lines have been fitted to the process, one before  $t = 0.6$  and one after. The ratio of the slopes is  $-0.12$  which makes us consider the regression coefficient  $\beta_1(t) = \beta_1 B_{0.6}(t)$  where  $B_{0.6}(t) = I(t \leq 0.6) - 0.12 I(t \geq 0.6)$ . Other piecewise constant regression coefficients  $\beta(t) = \beta_1 B_{t_0}(t)$  have been considered with changepoints at times  $t_0 \in \{0.45, 0.5, \dots, 0.7\}$ . For each time  $t_0$ , the ratio of slopes has been evaluated to determine the value which multiplies the coefficient in the second part on the study. The second covariate  $Z^2$ , however, seems to have a

constant regression coefficient since the process stands between the confidence bands and has a linear trend (Figure 5(b)). Therefore, we consider only the regression coefficient  $\beta_2(t) = \beta_2$ . Estimation results are given in Table 2. The proportional hazards model

$\beta_1(t)$	$\beta_1$	$\beta_1 B_{0.45}(t)$	$\beta_1 B_{0.5}(t)$	$\beta_1 B_{0.55}(t)$	$\beta_1 B_{0.6}(t)$	$\beta_1 B_{0.65}(t)$	$\beta_1 B_{0.7}(t)$
$\hat{\beta}_1$	0.45	0.93	0.96	0.89	0.95	0.86	0.72
$\hat{\beta}_2$	-0.73	-0.72	-0.73	-0.74	-0.79	-0.80	-0.77
$R^2$	0.24	0.35	0.37	0.35	0.39	0.37	0.32

Table 2: Maximum partial likelihood estimators  $\hat{\beta}(t)$  and  $R^2$  coefficients on a simulated dataset with  $\beta_1(t) = I(t \leq 0.5)$  and  $\beta_2(t) = -1$ .

gives an  $R^2$  of 0.24. The maximal  $R^2$  is obtained when considering  $\beta_1(t) = \beta_1 B_{0.6}(t)$ , with an increase of 60% compared to the proportional hazards model. Therefore, we choose the model with  $\beta_1(t) = \beta_1 B_{0.6}(t)$  and  $\beta_2(t) = \beta_2$ .

## 5 CLINICAL STUDY IN BREAST CANCER

We return to the motivating example of the 1504 patients suffering from breast cancer. These patients were followed over a period of 15 years at the Institut Curie in Paris, France. Several studies were based on these data. One sub-study considered the predictive effects of the prognostic factors; progesterone receptor status, the tumor size over 60 mm and the grading over 2. The multivariate standardized score process  $\hat{\Sigma}^{-1/2} \mathbf{U}^*(\mathbf{0}, \cdot)$  and its confidence band are plotted over time in Figure 6. In Figure 6(a), we illustrate the process corresponding to the tumor size effect. Clearly, the effect seems non-constant with slope gradually diminishing with time. So much so that the process ends up drifting beyond the limits of the 95% confidence band. A slightly more refined model providing a much better fit allows for a change in effect at time point  $t = 0.2$ . As in our simulated examples, two straight lines have been fitted to the curve before and after  $t = 0.2$ , leading us to consider the regression effect  $\beta_{size}(t) = \beta_0(I(t \leq 0.2) + 0.24I(t \geq 0.2))$ . From Table 3 we can quantify the predictive improvement of Model 2 (constant effects for hormone receptor status and grade, time dependent effects for tumor size) versus Model 1 (all 3 prognostic factors constant) by a greater than 30% increase in the size of  $R^2$ , from 0.29 to 0.39. Figure 6(b) represents the process for the effect of the progesterone receptor over time. Again there is some evidence of a changing slope, although much weaker than for tumor size and, indeed, the process remains within the limits of the confidence bands. We considered various potential regression effects: a changepoint model with a cut at time  $t = 0.5$ ,  $\beta_{rec0}(t) = \beta_0(I(t \leq 0.5) + 0.39I(t \geq 0.5))$  and several smooth parameters  $\beta_{rec1}(t) = \beta_0(1 - t)$ ,  $\beta_{rec2}(t) = \beta_0(1 - t)^2$ ,  $\beta_{rec3}(t) = \beta_0(1 - t^2)$  and  $\beta_{rec4}(t) = \beta_0 \log(t)$ . Figure 6(c) represents the process for the grading effect. There is a clear impression of the steepness of the negative slope attenuating with time. The

process reaches the limits of the confidence bands but does not go beyond them. The simpler model, i.e., proportional hazards effects implying a linear slope, may be good enough although, in a model building context, it is also worth considering one with time dependent effects. Specifically, we chose to also look at a model with piecewise constant coefficients  $\beta_{gra}(t) = \beta_0(t)(I(t \leq 0.4) + 0.69I(t \geq 0.4))$ .

All of these several combinations, alongside models with constant effects, were looked at. For each combination, the regression effects have been estimated by maximizing the partial likelihood and the  $R^2$  coefficient has been evaluated. Partial results are given in

Tumor size	Receptor	Grading	$R^2$
0.84	1.03	-0.68	0.29
$1.77(I(t \leq 0.2) + 0.24I(t \geq 0.2))$	1.03	-0.66	0.39
0.85	$-1.02 \log(t)$	-0.67	0.39
$1.74(I(t \leq 0.2) + 0.24I(t \geq 0.2))$	$-1.02 \log(t)$	-0.66	0.51
$1.72(I(t \leq 0.2) + 0.24I(t \geq 0.2))$	$-1.02 \log(t)$	$-0.82I(t \leq 0.4) + 0.69I(t \geq 0.4)$	0.52

Table 3: Maximum partial likelihood estimators and  $R^2$  coefficients on the breast cancer dataset.

Table 3. The proportional hazards model gives an  $R^2$  coefficient of 0.29. As mentioned above, a more involved model allowing for the effect of tumor size to assume a simple time dependency results in a big jump in observed predictability of an order greater than 30%. The highest  $R^2$  is obtained with changepoints for tumor size and grading covariates, with a function of  $\log(t)$  for the effect of progesterone receptor. The predictive accuracy of this model has increased by 80% compared to the predictive accuracy of the corresponding proportional hazards model. This gives a strong indication that, as far as prediction is concerned, significant improvement can be consequent on allowing time dependency. On the other hand, allowing for time dependency grade, having already accounted for the joint effects of tumor size and receptor status, results in an increase in  $R^2$  from 0.51 to 0.52. Such a negligible increase does not justify the added complexity of the model so that, provided the other two risk factors are included, it makes sense to restrict the effects of grade to be constant.

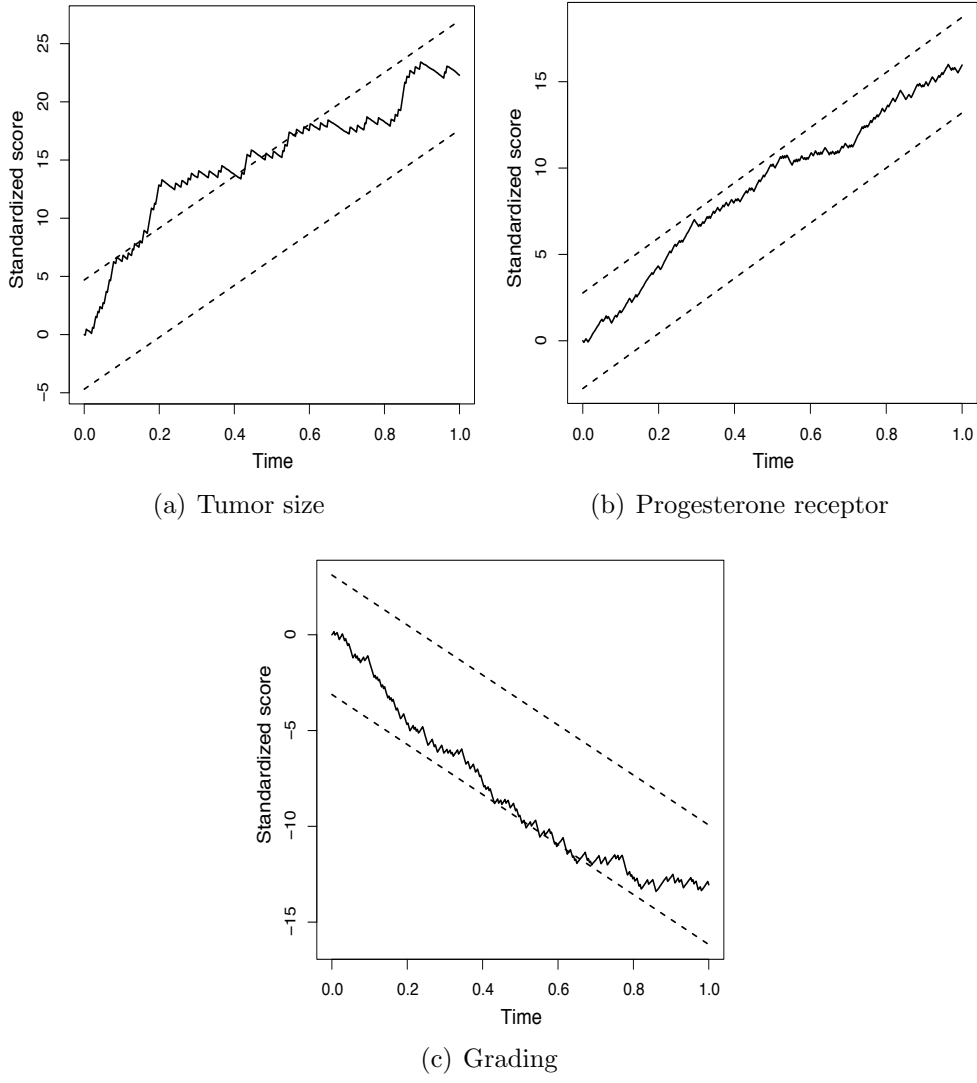


Figure 6: Standardized score process  $\hat{\Sigma}^{-1}\mathbf{U}^*(0, \cdot)$  (solid lines) and its confidence bands (dashed lines) on the breastcancer dataset for tumor size, progesterone receptor and grading.

## 6 DISCUSSION

The related and complementary techniques of goodness of fit and predictive ability provide a coherent way to construct models. Intuitively, models constructed in this way ought provide a better predictive performance. This intuition is correct and is supported by the theoretical results of this paper. Our preference is to appeal to techniques based on the Schoenfeld residual processes for proportional and non-proportional hazards models since these processes provide the basis for both of these techniques. A large number of competing approaches appears possible since there is a large body of literature on goodness of fit procedures and a large body on predictive measures. Combinations of these could provide tools analogous to those described here. However, in order to make analogous claims to ours concerning predictive performance for some particular combination, we would require equivalent theorems to those presented in Sections 2 and 3.

We might consider that the first step away from a proportional hazards model is a similar model but with a changepoint. Before the changepoint we have one particular proportional hazards model whereas, after the changepoint, we have a model with a different value of  $\beta$ . The methods described here would enable us to estimate the changepoint itself as well as the values of  $\beta$ , before and after the changepoint. Extending this to more than a single changepoint is, at least in theory, straightforward. This suggests one possible systematic way of model construction. Another extension that would be worth considering is the estimation of the process drift with non-parametric estimation techniques in order to estimate the cumulative regression effect  $\int_0^t \beta(s)ds$ .

## A Proof of Theorem 1

Define the filtration  $\{\mathcal{F}_t\}_{t \in [0,1]} = \sigma\{\bar{N}_j(u), \bar{Y}_j(u^+), \mathbf{Z}_j(u^+), j = 1, \dots, n, 0 \leq u \leq t\}$ . Each failure time  $t_i$  is a  $\mathcal{F}_t$ -stopping time. Consider the conditional expectation

$$\mathbf{E}_{\beta(t)}(\mathbf{h}|\mathcal{F}_t) = \sum_{j=1}^n \mathbf{h}_j(t) \pi_j(\beta(t), t),$$

where  $\mathbf{h}_j$  is a  $\mathbb{R}$  or  $\mathbb{R}^p$ -predictable process for individual  $j$ . In order to simplify the notation, denote  $\mathbf{E}_{\beta(t)}(\mathbf{h}|t) = \mathbf{E}_{\beta(t)}(\mathbf{h}|\mathcal{F}_t)$  and  $\mathbf{V}_{\beta(t)}(\mathbf{h}|t) = \mathbf{E}_{\beta(t)}(\mathbf{h}^{\otimes 2}|\mathcal{F}_t) - \mathbf{E}_{\beta(t)}(\mathbf{h}|\mathcal{F}_t)^{\otimes 2}$ . The first part of the proof shows the convergence in distribution of  $\mathbf{U}^*(\beta_0, \cdot) - \sqrt{k_n} \mathbf{C}_n$  to a multivariate Wiener process as  $n$  increases without bound. Denote  $\mathbf{X}_n$  the right-continuous with left-hand limits process  $\mathbf{X}_n$ , with a jump at each  $t_i$  such that

$$\mathbf{X}_n(t) = \frac{1}{\sqrt{k_n}} \sum_{i=1}^{\lfloor tk_n \rfloor} \mathbf{V}_{\beta_0}(Z|t_i)^{-1/2} \{ \mathcal{Z}(t_i) - \mathbf{E}_{\beta(t_i)}(Z|t_i) \}, \quad 0 \leq t \leq 1.$$

Notice that  $\mathbf{X}_n(t_i) = \mathbf{U}^*(\beta_0, t_i) - \sqrt{k_n} \mathbf{C}_n(t_i)$  at each  $t_i = i/k_n$ ,  $i = 1, \dots, k_n$ . Denote  $\boldsymbol{\xi}_{i,k_n} = (\xi_{i,k_n}^1, \dots, \xi_{i,k_n}^p)$  the  $i$ th  $\mathbb{R}^p$ -valued increment of the process  $\mathbf{X}_n$ . Notice that  $\boldsymbol{\xi}_{i,k_n}$

is  $\mathcal{F}_{t_i}$ -measurable. Then,

$$\left\| \mathbf{U}^*(\beta_0, \cdot) - \sqrt{k_n} \mathbf{C}_n - \mathbf{W}_p \right\| \leq \left\| \mathbf{U}^*(\beta_0, \cdot) - \sqrt{k_n} \mathbf{C}_n - \mathbf{X}_n \right\| + \left\| \mathbf{X}_n - \mathbf{W}_p \right\|.$$

The first term on the right hand side converges to 0 as  $n$  increases without bound by the existence of a moment of order 3 of the increments  $\xi_{i,k_n}$ . The convergence in distribution of  $\mathbf{X}_n$  to  $\mathbf{W}_p$  is given by the multivariate functional central limit theorem of Helland (1982) of which hypotheses are verified in Supplementary Material. It remains to prove equation (7). A multidimensional Taylor-Lagrange series expansion gives

$$\left\| \mathbf{E}_{\beta(t)}(Z|t) - \mathbf{E}_{\beta_0}(Z|t) - \mathbf{V}_{\beta_0}(Z|t) \{\beta(t) - \beta_0\} \right\| \leq \frac{M_n}{2} \|\beta(t) - \beta_0\|^2.$$

Therefore,

$$\begin{aligned} & \left\| C_n(t) - \Sigma^{1/2} \int_0^t \{\beta(s) - \beta_0\} ds \right\| \\ & \leq \frac{1}{k_n} \sum_{i=1}^{\lfloor tk_n \rfloor} \left\| \mathbf{V}_{\beta_0}(Z|t_i)^{-1/2} \{ \mathbf{E}_{\beta_0}(Z|t_i) - \mathbf{E}_{\beta(t_i)}(Z|t_i) - \mathbf{V}_{\beta_0}(Z|t_i) \{\beta(t_i) - \beta_0\} \} \right\| \\ & \quad + \frac{1}{k_n} \sum_{i=1}^{\lfloor tk_n \rfloor} \left\| \left( \mathbf{V}_{\beta_0}(Z|t_i)^{1/2} - \Sigma^{1/2} \right) \{\beta(t_i) - \beta_0\} \right\| \\ & \quad + \left\| \Sigma^{1/2} \left( \frac{1}{k_n} \sum_{i=1}^{\lfloor tk_n \rfloor} \beta(t_i) - \int_0^t \beta(s) ds + \beta_0 \left( \frac{\lfloor tk_n \rfloor}{k_n} - t \right) \right) \right\| \\ & \leq \frac{pM_n}{2k_n} \sum_{i=1}^{\lfloor tk_n \rfloor} \left\| \mathbf{V}_{\beta_0}(Z|t_i)^{-1/2} \right\| \|\beta(t_i) - \beta_0\|^2 + \frac{p}{k_n} \sum_{i=1}^{\lfloor tk_n \rfloor} \left\| \mathbf{V}_{\beta_0}(Z|t_i)^{1/2} - \Sigma^{1/2} \right\| \|\beta(t_i) - \beta_0\| \\ & \quad + p \left\| \Sigma^{1/2} \right\| \left\| \frac{1}{k_n} \sum_{i=1}^{\lfloor tk_n \rfloor} \beta(t_i) - \int_0^t \beta(s) ds \right\| + \left| \frac{\lfloor tk_n \rfloor}{k_n} - t \right| \left\| \Sigma^{1/2} \beta_0 \right\| \\ & \leq \frac{\lfloor tk_n \rfloor}{k_n} \frac{pM_n}{2} \max_{i=1, \dots, \lfloor tk_n \rfloor} \left\| \mathbf{V}_{\beta_0}(Z|t_i)^{-1/2} \right\| \max_{i=1, \dots, \lfloor tk_n \rfloor} \|\beta(t_i) - \beta_0\|^2 \\ & \quad + p \frac{\lfloor tk_n \rfloor}{k_n} \max_{i=1, \dots, \lfloor tk_n \rfloor} \|\beta(t_i) - \beta_0\| \max_{i=1, \dots, \lfloor tk_n \rfloor} \left\| \mathbf{V}_{\beta_0}(Z|t_i)^{1/2} - \Sigma^{1/2} \right\| \\ & \quad + p \left\| \Sigma^{1/2} \right\| \max_{l=1, \dots, p} \left| \frac{1}{k_n} \sum_{i=1}^{\lfloor tk_n \rfloor} \beta(t_i)_l - \int_0^t \beta(s)_l ds \right| + p \left| \frac{\lfloor tk_n \rfloor}{k_n} - t \right| \left\| \Sigma^{1/2} \right\| \|\beta_0\|. \end{aligned}$$

This norm converges to 0 in probability as  $n$  increases without bound by the boundedness of the variances (Assumptions A and B), their convergence to  $\Sigma$  (Assumption C) and the convergence to 0 of  $M_n$  as  $n \rightarrow \infty$ .  $\square$



## B Proof of Proposition 1

Let  $t \in [0, 1]$ . By Theorem 1 and since  $\hat{\Sigma}^{-1/2}$  is a consistent estimator of  $\Sigma^{-1/2}$ , in addition to Slutsky's lemma, we have  $\hat{\Sigma}^{-1/2} (\mathbf{U}^*(\beta_0, t) - t\mathbf{U}^*(\beta_0, 1)) \xrightarrow[n \rightarrow +\infty]{D} \Sigma^{-1/2} \mathbf{B}_p(t)$ . Therefore,

$$\left\| \hat{\Sigma}_{\cdot, i}^{-1/2} \right\|_2^{-1} \hat{\Sigma}^{-1/2} (\mathbf{U}^*(\beta_0, t) - t\mathbf{U}^*(\beta_0, 1)) \xrightarrow[n \rightarrow +\infty]{D} B(t),$$

where  $B$  is a Brownian Bridge. The result follow from the knowledge of the limit distribution of the supremum of the absolute value of a Brownian bridge, which is the Kolmogorov distribution.  $\square$

## C Proof of Theorem 2

We consider first the univariate case, in which  $p = 1$ . Let us study the numerator of the  $R^2$  coefficient defined in equation (12). We have

$$\begin{aligned} & \frac{1}{k_n} \sum_{i=1}^{k_n} (\mathcal{Z}(t_i) - \mathbf{E}_{\alpha(t_i)}(Z \mid t_i))^2 \\ &= \frac{1}{k_n} \sum_{i=1}^{k_n} (\mathcal{Z}(t_i) - \mathbf{E}_{\beta(t_i)}(Z \mid t_i))^2 + \frac{1}{k_n} \sum_{i=1}^{k_n} (\mathbf{E}_{\beta(t_i)}(Z \mid t_i) - \mathbf{E}_{\alpha(t_i)}(Z \mid t_i))^2 \\ &+ \frac{2}{k_n} \sum_{i=1}^{k_n} (\mathcal{Z}(t_i) - \mathbf{E}_{\beta(t_i)}(Z \mid t_i)) (\mathbf{E}_{\beta(t_i)}(Z \mid t_i) - \mathbf{E}_{\alpha(t_i)}(Z \mid t_i)). \end{aligned} \quad (13)$$

Let us study the right-hand side of equation (13). Recall that the random variables  $\mathcal{Z}(t_i) - \mathbf{E}_{\beta(t_i)}(Z \mid t_i)$  are independent for  $i = 1, \dots, k_n$ , that  $Z(t)$  admits a moment of order 4 and  $E \left( (\mathcal{Z}(t_i) - \mathbf{E}_{\beta(t_i)}(Z \mid t_i))^2 \right) = E \left( \mathbf{V}_{\beta(t_i)}(Z \mid t_i) \right)$ . Therefore,

$$\lim_{n \rightarrow \infty} \sum_{i=1}^{k_n} \frac{1}{i^2} E \left( \{ \mathcal{Z}(t_i) - \mathbf{E}_{\beta(t_i)}(Z \mid t_i) \}^4 \right) < \infty.$$

Markov's law of large numbers for independent and non-identically distributed random variables imply that

$$\frac{1}{k_n} \sum_{i=1}^{k_n} (\mathcal{Z}(t_i) - \mathbf{E}_{\beta(t_i)}(Z \mid t_i))^2 - \frac{1}{k_n} \sum_{i=1}^{k_n} E \left( \mathbf{V}_{\beta(t_i)}(Z \mid t_i) \right) \xrightarrow[n \rightarrow \infty]{\mathbb{P}} 0. \quad (14)$$

By Lemma 1 of Chauvel and O'Quigley (2014), we have

$$\frac{1}{k_n} \sum_{i=1}^{k_n} \mathbf{V}_{\beta(t_i)}(Z \mid t_i) \xrightarrow[n \rightarrow \infty]{\mathbb{P}} \int_0^1 v(\beta(t), t) dt. \quad (15)$$

Conditional empirical variances are almost surely bounded implying that

$$\frac{1}{k_n} \sum_{i=1}^{k_n} (\mathcal{Z}(t_i) - \mathbf{E}_{\beta(t_i)}(\mathcal{Z} \mid t_i))^2 \xrightarrow[n \rightarrow \infty]{\mathbb{P}} \int_0^1 v(\beta(t), t) dt.$$

The convergence of the second term of equation (13) is again obtained by Lemma 1 of Chauvel and O’Quigley (2014):

$$\frac{1}{k_n} \sum_{i=1}^{k_n} (\mathbf{E}_{\beta(t_i)}(\mathcal{Z} \mid t_i) - \mathbf{E}_{\alpha(t_i)}(\mathcal{Z} \mid t_i))^2 \xrightarrow[n \rightarrow \infty]{\mathbb{P}} \int_0^1 (e(\alpha(t), t) - e(\beta(t), t))^2 dt.$$

Finally, the last term of equation (13) converges in probability to 0 when  $n \rightarrow \infty$  by Markov’s law of large numbers. Thus,

$$\lim_{n \rightarrow \infty} R^2(\alpha(t)) = 1 - \frac{\int_0^1 v(\beta(t), t) dt + \int_0^1 (e(\alpha(t), t) - e(\beta(t), t))^2 dt}{\int_0^1 v(\beta(t), t) dt + \int_0^1 (e(0, t) - e(\beta(t), t))^2 dt}, \quad (16)$$

and  $\lim_{n \rightarrow \infty} R^2$  reaches its maximum in  $\beta(t)$ .

For the multivariate case ( $p > 1$ ), similar arguments lead to the limit

$$\lim_{n \rightarrow \infty} R^2(\alpha(t)) = 1 - \frac{\int_0^1 \alpha(t)^T v(\beta(t), t) \alpha(t) dt + \int_0^1 (\alpha(t)^T \{e(\beta(t), t) - e(\alpha(t), t)\})^2 dt}{\int_0^1 \alpha(t)^T v(\beta(t), t) \alpha(t) dt + \int_0^1 (\alpha(t)^T \{e(\beta(t), t) - e(0, t)\})^2 dt}. \quad (17)$$

Finally,  $\left| R^2(\beta(t)) - R^2(\hat{\beta}(t)) \right| \xrightarrow[n \rightarrow \infty]{a.s.} 0$ , as  $\hat{\beta}(t)$  is a consistent estimator of  $\beta(t)$ .  $\square$

## D Proof of Theorem 3

Let  $p = 1$ ,  $\alpha(t) \in \mathcal{B}$  and assume that A and B are verified. A Taylor series expansion of  $e(\alpha(t), t)$  in equation (16) gives

$$\lim_{n \rightarrow \infty} R^2(\alpha(t)) = 1 - \frac{\int_0^1 v(\beta(t), t) dt + \int_0^1 (\alpha(t) - \beta(t))^2 v(c(t), t) dt}{\int_0^1 v(\beta(t), t) dt + \int_0^1 (e(0, t) - e(\beta(t), t))^2 dt},$$

where  $c(t)$  lies between  $\alpha(t)$  and  $\beta(t)$ . Therefore, minimizing  $\lim_{n \rightarrow \infty} R^2(\alpha(t))$  in  $\alpha(t)$  reduces to minimize  $\int_0^1 (\alpha(t) - \beta(t))^2 v(c(t), t) dt$ .  $\square$

## Supplementary Material

Consider the setting of the proof of Theorem 1. Let us verify that the hypotheses of the functional central limit theorem for martingale differences of Helland (1982) are satisfied.

Let  $t \in [0, 1]$  and  $l, m = 1, \dots, p$ , with  $l \neq m$ . Denote  $\mathbf{e}_l$  the  $l$ th vector of the standard basis of  $\mathbb{R}^p$ : all of its elements are null except for its  $l$ th element which equals 1. Then,

$$\xi_{i,k_n}^l = \mathbf{e}_l^T \boldsymbol{\xi}_{i,k_n} = \boldsymbol{\xi}_{i,k_n}^T \mathbf{e}_l \in \mathbb{R}.$$

A. (Martingale difference array.) Using the inclusions of the  $\sigma$ -algebras  $\mathcal{F}_{t_{i-1}} \subset \mathcal{F}_{t_i}$  and the centering of the increments, we have

$$E_{\beta(t_{i-1})}(\xi_{i,k_n}^l | t_{i-1}) = E_{\beta(t_{i-1})}(E_{\beta(t_i)}(\xi_{i,k_n}^l | t_i) | t_{i-1}) = 0.$$

B. (Uncorrelatedness.) Notice that

$$\begin{aligned} E_{\beta(t_i)}(\xi_{i,k_n}^l \xi_{i,k_n}^m | t_i) &= \mathbf{e}_l^T \mathbf{E}_{\beta(t_i)}(\boldsymbol{\xi}_{i,k_n} \boldsymbol{\xi}_{i,k_n}^T | t_i) \mathbf{e}_m \\ &= \frac{1}{k_n} \mathbf{e}_l^T \mathbf{V}_{\beta_0}(Z|t_i)^{-1/2} \mathbf{V}_{\beta(t_i)}(Z|t_i) \mathbf{V}_{\beta_0}(Z|t_i)^{-1/2} \mathbf{e}_m \end{aligned}$$

Therefore, using the inclusion of sigma-algebras,

$$\begin{aligned} &E \left( \left| \sum_{i=1}^{\lfloor tk_n \rfloor} E_{\beta(t_{i-1})}(\xi_{i,k_n}^l \xi_{i,k_n}^m | t_{i-1}) \right| \right) \\ &= E \left( \left| \frac{1}{k_n} \sum_{i=1}^{\lfloor tk_n \rfloor} E_{\beta(t_{i-1})}(\mathbf{e}_l^T \mathbf{V}_{\beta_0}(Z|t_i)^{-1/2} \mathbf{V}_{\beta(t_i)}(Z|t_i) \mathbf{V}_{\beta_0}(Z|t_i)^{-1/2} \mathbf{e}_m | t_{i-1}) \right| \right) \\ &\leq \frac{1}{k_n} \sum_{i=1}^{\lfloor tk_n \rfloor} E(|\mathbf{e}_l^T \mathbf{V}_{\beta_0}(Z|t_i)^{-1/2} \mathbf{V}_{\beta(t_i)}(Z|t_i) \mathbf{V}_{\beta_0}(Z|t_i)^{-1/2} \mathbf{e}_m|) \\ &\leq \frac{\lfloor tk_n \rfloor}{k_n} \max_{i=1, \dots, \lfloor k_n t \rfloor} E(|\mathbf{e}_l^T \mathbf{V}_{\beta_0}(Z|t_i)^{-1/2} \mathbf{V}_{\beta(t_i)}(Z|t_i) \mathbf{V}_{\beta_0}(Z|t_i)^{-1/2} \mathbf{e}_m|). \end{aligned}$$

By assumption C and the continuous mapping theorem for matrices and vectors,

$$\mathbf{e}_l^T \mathbf{V}_{\beta_0}(Z|t_i)^{-1/2} \mathbf{V}_{\beta(t_i)}(Z|t_i) \mathbf{V}_{\beta_0}(Z|t_i)^{-1/2} \mathbf{e}_m \xrightarrow[n \rightarrow \infty]{P} \mathbf{e}_l^T \boldsymbol{\Sigma}^{-1/2} \boldsymbol{\Sigma} \boldsymbol{\Sigma}^{-1/2} \mathbf{e}_m = 0.$$

This convergence is also a convergence in mean by the almost sure boundedness of each quantity. Thus,  $E_{\beta(t_{i-1})}(\xi_{i,k_n}^l \xi_{i,k_n}^m | t_{i-1}) \xrightarrow[n \rightarrow \infty]{L^1} 0$ .

C. (Variance.) Denote  $\mathbf{I}_p$  the identity matrix of dimension  $p \times p$ . The same arguments leads us to the following equality

$$\begin{aligned} &\sum_{i=1}^{\lfloor tk_n \rfloor} E_{\beta(t_{i-1})} \left( \left( \xi_{i,k_n}^l \right)^2 | t_{i-1} \right) - t \\ &= \frac{1}{k_n} \sum_{i=1}^{\lfloor tk_n \rfloor} E_{\beta(t_{i-1})}(\mathbf{e}_l^T \{ \mathbf{V}_{\beta_0}(Z|t_i)^{-1/2} \mathbf{V}_{\beta(t_i)}(Z|t_i) \mathbf{V}_{\beta_0}(Z|t_i)^{-1/2} - \mathbf{I}_p \} \mathbf{e}_l | t_{i-1}) + \frac{\lfloor k_n t \rfloor}{k_n} - t. \end{aligned}$$

Thus,

$$\begin{aligned} & E \left( \left| \sum_{i=1}^{\lfloor tk_n \rfloor} E_{\beta(t_{i-1})} \left( (\xi_{i,k_n}^l)^2 \middle| t_{i-1} \right) - t \right| \right) \\ & \leq \frac{1}{k_n} \sum_{i=1}^{\lfloor tk_n \rfloor} E \left( \left| \mathbf{e}_l^T \{ \mathbf{V}_{\beta_0}(Z|t_i)^{-1/2} \mathbf{V}_{\beta(t_i)}(Z|t_i) \mathbf{V}_{\beta_0}(Z|t_i)^{-1/2} - \mathbf{I}_p \} \mathbf{e}_l \right| \right) + \left| \frac{\lfloor k_n t \rfloor}{k_n} - t \right|. \end{aligned}$$

Again, assumption C , the continuous mapping theorem and the almost sure boundedness of the variances imply

$$\mathbf{e}_l^T \{ \mathbf{V}_{\beta_0}(Z|t_i)^{-1/2} \mathbf{V}_{\beta(t_i)}(Z|t_i) \mathbf{V}_{\beta_0}(Z|t_i)^{-1/2} - \mathbf{I}_p \} \mathbf{e}_l \xrightarrow[n \rightarrow \infty]{L^1} \mathbf{e}_l^T \{ \Sigma^{-1/2} \Sigma \Sigma^{-1/2} - \mathbf{I} \} \mathbf{e}_l = 0.$$

$$\text{Therefore, } \sum_{i=1}^{\lfloor tk_n \rfloor} E_{\beta(t_{i-1})} \left( (\xi_{i,k_n}^l)^2 \middle| t_{i-1} \right) \xrightarrow[n \rightarrow \infty]{L^1} t.$$

D. (Lyapunov condition.) By the boundedness of the increments  $(\xi_{i,k_n}^l)_i$ , there exists a constant  $C \in [0, +\infty[$  such that for all  $i = 1, \dots, k_n$ ,  $E_{\beta(t_i)} \left( |\xi_{i,k_n}^l|^3 \middle| t_i \right) \leq C$  almost surely. Thus,

$$\sum_{i=1}^{\lfloor k_n t \rfloor} E_{\beta(t_{i-1})} \left( |\xi_{i,k_n}^l|^3 \middle| t_{i-1} \right) \leq \frac{1}{k_n^{3/2}} \sum_{i=1}^{\lfloor k_n t \rfloor} E_{\beta(t_{i-1})} \left( E_{\beta(t_i)} \left( |\xi_{i,k_n}^l|^3 \middle| t_i \right) \middle| t_{i-1} \right) \leq \frac{\lfloor k_n t \rfloor}{k_n^{3/2}} C.$$

$$\text{Hence, } \sum_{i=1}^{\lfloor k_n t \rfloor} E_{\beta(t_{i-1})} \left( |\xi_{i,k_n}^l|^3 \middle| t_{i-1} \right) \xrightarrow[n \rightarrow \infty]{P} 0.$$

As a conclusion, all hypotheses of Helland's multivariate functional central limit theorem are gathered and  $\mathbf{X}_n$  converges weakly to a multivariate Wiener process  $\mathbf{W}_p$  as  $n$  increases without bound.

## References

- Andersen, P. (1982). Testing goodness of fit of Cox's regression and life model. *Biometrics*, 38:67–77.
- Andersen, P. and Gill, R. (1982). Cox's regression model for counting processes: A large sample study. *The Annals of Statistics*, 10:1100–1120.
- Arjas, E. (1988). A graphical method for assessing goodness of fit in Cox's proportional hazards model. *Journal of the American Statistical Association*, 83(401):204–212.
- Barlow, W. E. and Prentice, R. L. (1988). Residuals for relative risk regression. *Biometrika*, 75(1):65–74.

- Cai, Z. and Sun, Y. (2003). Local linear estimation for time-dependent coefficients in Cox's regression models. *Scandinavian Journal of Statistics*, 30:93–111.
- Chauvel, C. and O'Quigley, J. (2014). Tests for comparing estimated survival functions. *Biometrika*, doi: 10.1093/biomet/asu015.
- Choodari-Oskoei, B., Royston, P., and Parmar, M. K. B. (2012). A simulation study of predictive ability measures in a survival model I: Explained variation measures. *Statistics in Medicine*, 31(23):2627–2643.
- Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society. Series B*, 34(2):187–220.
- Cox, D. R. (1975). Partial likelihood. *Biometrika*, 63:269–276.
- Grambsch, P. M. and Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, (3):515–526.
- Hastie, T. and Tibshirani, R. (1990). Exploring the nature of covariate effects in the proportional hazards model. *Biometrics*, 46:1005–1016.
- Helland, I. (1982). Central limit theorems for martingales with discrete or continuous time. *Scandinavian Journal of Statistics*, 9:79–94.
- Hielscher, T., Zucknick, M., Werft, W., and Benner, A. (2010). On the prognostic value of survival models with application to gene expression signatures. *Statistics in Medicine*, 29(7–8):818–829.
- Kay, R. (1977). Proportional hazard regression models and the analysis of censored survival data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 26(3):227–237.
- Khmaladze, E. V. (1981). Martingale approach in the theory of goodness-of-fit tests. *Theory of Probability and its applications*, 26(2):240–257.
- Klein, J. and Moeschberger, M. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*. Springer.
- Lin, D., Robins, J., and Wei, L. (1996). Comparing two failure time distributions in the presence of dependent censoring. *Biometrika*, 83:381–393.
- Lin, D., Wei, L., and Ying, Z. (1993). Checking the Cox model with cumulative sums of martingale based residuals. *Biometrika*, 80:557–572.
- Martinussen, T. and Scheike, T. (2005). *Dynamic Regression Models for Survival Data*. Springer.
- Müller, M., Döring, A., Küchenhoff, H., Lamina, C., Malzahn, D., Bickeböller, H., Vollmert, C., Klopp, N., Meisinger, C., Heinrich, J., Kronenberg, F., Erich Wichmann, H., and Heid, I. (2008). Quantifying the contribution of genetic variants for survival phenotypes. *Genetic Epidemiology*, 32(6):574–585.

- Murhpy, S. and Sen, P. (1991). Time-dependent coefficients in a Cox-type regression model. *Stochastic Processes and their Applications*, 39:153–180.
- O’Quigley, J. (2003). Khmaladze-type graphical evaluation of the proportional hazards assumption. *Biometrika*, 90:577–584.
- O’Quigley, J. (2008). *Proportional Hazards Regression*. Springer. New York.
- O’Quigley, J. and Flandre, P. (1994). Predictive capability of proportional hazards regression. *Proceedings of the National Academy of Sciences*, 91(6):2310–2314.
- O’Quigley, J. and Xu, R. (2012). Explained variation in proportional hazards regression. In Crowley, J. and Hoering, A., editors, *Handbook of Statistics in Clinical Oncology, Third Edition*, pages 487–504. Chapman and Hall, CRC.
- Sasieni, P. and Winnett, A. (2003). Martingale difference residuals as a diagnostic tool for the Cox model. *Biometrika*, 90:899–912.
- Scheike, T. and Martinussen, T. (2004). Maximum likelihood estimation for Cox’s regression model under casecohort sampling. *Scandinavian Journal of Statistics*, 31(2):283–293.
- Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, 69:239–241.
- Therneau, T. and Grambsch, P. (2000). *Modeling Survival Data: Extending the Cox Model*. Springer. New York.
- Therneau, T., Grambsch, P., and Fleming, T. (1990). Martingale-based residuals for survival models. *Biometrika*, 77(1):147–160.
- Wei, L. (1984). Testing goodness-of-fit for proportional hazards model with censored observations. *Journal of the American Statistical Association*, 79:649–652.
- Winnett, A. and Sasieni, P. (2003). Iterated residuals and time-varying covariate effects in Cox regression. *Journal of the Royal Statistical Society. Series B*, 65:473–488.
- Zucker, D. M. and Lakatos, E. (1990). Weighted log rank type statistics for comparing survival curves when there is a time lag in the effectiveness of treatment. *Biometrika*, 77:853–864.